AD_______________

Award Number: MIPR OEC5E50082

TITLE: The Application of Information Mining Technology to the Total Army Injury and Health Outcomes Database (TAIHOD)

PRINCIPAL INVESTIGATOR: LTC Paul Amoroso

CONTRACTING ORGANIZATION: U.S. Army Research Institute of
Environmental Medicine
Natick, Massachusetts 01760-5007

REPORT DATE: May 2000

TYPE OF REPORT: Midterm

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
May 2000

3. REPORT TYPE AND DATES COVERED
Midterm (15 Feb 00 - 30 Apr 00)

4. TITLE AND SUBTITLE
The Application of Information Mining Technology to the Total Army Injury and Health Outcomes Database (TAIHOD)

5. FUNDING NUMBERS
MIPR0EC5E50082

6. AUTHOR(S)
LTC Paul Amoroso

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Georgetown University
Washington, DC 20057

E-Mail: paul.amoroso@det.amedd.army.mil

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

14. SUBJECT TERMS

15. NUMBER OF PAGES
5

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	Unlimited

MidTerm Overall Evaluation Report

PROPOSAL: 1999000212

TITLE: The Application of Information Mining Technology to the Total Army Injury and Health Outcomes Database (TAIHOD)

1. ACCOMPLISHMENTS:

The principal accomplishments for the first half of this grant period were in training staff, procurement of equipment and software, and execution of contracts. Considerable progress, including many preliminary analyses, was also made in developing a strategy for accomplishing the objectives of the text mining portion of the proposal. We also procured several important new data sets directly relevant to the proposal. STAFF TRAINING: A 60-training day educational package was also purchased from SAS Institute Inc. This educational package allows the TAIHOD staff to take 60 days of training classes from SAS Institute in order to learn and exploit the data mining and warehouse administration software purchased for this grant as well as several related statistical procedures. The following training is completed or in process: 1. SAS Enterprise Miner, April 17-19, Dallas, TX. Four individuals. This class presented an introduction to using Enterprise Miner and to the application of data mining techniques. 2. SAS Warehouse Administrator, April 25-28, Boston, MA. Four individuals. This class presented an introduction to using Warehouse Administrator and developing a data warehouse. 3. Introduction to Time Series Forecasting using SAS/ETS Software, June 26-28, New York, NY, One individual. This course uses the time series forecasting system and the SAS procedures ARIMA, FORECAST and EXPAND. 4. Advanced General Linear Models with an Emphasis on Mixed Models, July 31-August 2, New York, NY, One individual. This course used the SAS procedures GLM and MIXED to estimate variance components and produce appropriate test statistics for fixed effects. 5. JMP Software: Statistical Data Exploration, August 29, Boston, MA, One individual. This course presented an introduction to JMP Software for visualization of data. 6. SQL Processing with the SAS System, September 27-28, Boston, MA. One individual. This course uses the SQL procedure to produce reports, data sets and views and to update data. 7. The principal investigator and one of the UMA collaborators attended a local conference, KDD-2000: The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining August 20-23, 2000, Boston, MA. 8. Two individuals attended the M2000 - Data Mining Technology Conference October 3-4, 2000 at SAS World Headquarters Cary, North Carolina. 9. One individual is attending a Cisco training course on switches, hubs, and network administration. PROCUREMENT OF EQUIPMENT AND SOFTWARE: Many purchases were made, including the core software for the project, a new database server, a hub and switch, and a color printer for exploitation of data visualization techniques. The two software products purchased were SAS Warehouse Administrator (version 2.0) and SAS Enterprise Miner (version 4.0). It took several months for the contract to be finalized, and we did not actually receive the software until June 2000. In addition, while not representing a software purchase per se, a contract with the University of Massachusetts Computer Science Department is expected to produce a customized version of a software product called Proximity. We are also currently in discussions with SAS Inc regarding a possible partnership in alpha testing of a new feature of their datamining softare that will allow text mining to be integrated into version 4.1. A new database server (HP LH6000R) that is well suited to handling the particular structure and analytic design of the TAIHOD database, and the processing demands of Enterprise Miner, has been ordered. Portions of the server have already been received; delivery of the remaining components is anticipated within 14 days. Configuration, testing, and transfer of data will likely take several additional weeks. EXECUTION OF CONTRACTS: While it took several months to finalize, we have obtained membership in the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts Amherst (UMA) Computer Science Department. Membership in the CIIR was a necessary condition of our collaboration with the computer science department at UMA. Collaborators at UMA are continuing to develop and customize a software package for us called Proximity, and are also developing a program to "anonymize" data by removing names from a long text string. In conjunction with this contract, we have hired a student contractor to act as a liaison between the CIIR and TAIHOD, and to assist in the creation of a database with text fields for the CIIR to analyze in accordance with their statement of work. Additionally, a separate contract directly with the UMA computer science department has finally been completed. This contract could not be executed until early September and was administratively time consuming to complete. However, now that it is in place, work is expected to progress in a timely fashion. PRELIMINARY ANALYSES AND RESEARCH PLAN FOR TEXT MINING: Our UMA collaborators will attempt to replicate a portion of a previously published study (Amoroso PJ, Bell NS, Jones BH. Injury among female and male Army parachutists. Aviat Space Environ Med. 1997; 68(11):1006-1011.) as a pilot test of the Proximity software product. We will use a similar dataset of parachute injuries identified by narratives from the Safety Center, and use Proximity software to first replicate the previous work and then look for new relationships in the data through mining the narrative text. We have subset and sanitized a sample of 560 parachute injury cases (removed names and other identifiers), coded them as to type of parachute injury, and sent them to UMass for analysis. They will use this sample to test the software, then use the entire dataset of 5,600 parachuting injury cases for additional analysis. After accomplishing this goal, we will experiment with using motor vehicle accident narratives (available in far greater numbers) in a similar fashion according to the UMA/CS statement of work. PROCUREMENT OF ADDITIONAL DATA: After many months of negotiation, we have reached an agreement with OTSG, PASBA, and TMSSC to obtain "injury comment fields" from all Army injury hospitalizations since 1990. This data is stored on CHCS computers at individual MTFs and must be extracted site-by-site and forwarded to us through TMSSC. This data will have significant bearing on the text mining software development as well as the two deliverables promised in the proposal on Gulf War illness and motor vehicle injuries. We also obtained data on family violence from the Army Central Registry. This data will be used in the study of deployment related outcomes and may have utility in the development of the deliverable related to Gulf War Illness. An additional dataset containing surveys of active duty soldiers has been cleared for transfer to us from DMDC. This data will provide important job satisfaction, family economics, and other information important to the study of deployment related conditions and disability.

2. PROBLEMS:

1. Problems/Issues with Software: The contract negotiations with SAS Institute Inc. were complicated and time consuming. While negotiations began in February, we did not receive the software until June. Therefore, we could not even begin to build the data warehouse until June. Full capability of the data warehousing and data mining tools is still several months away. On the other hand, we were able to negotiate a reduced price for the software. This factor has resulted in an opportunity to obtain additional consulting services, to upgrade our database server, and to ultimately speed up implementation of and improve the quality of the data warehouse. Once the warehouse is fully implemented, we should be in a more favorable position to exploit it. Furthermore, the structure of the contract places us in a favorable financial position going forward such that we are more likely to be able to continue the use of these tools in FY01 and FY02. 2. Problems/issues with the UMA/CS contracts: Our planned collaboration with the University of Massachusetts computer science department required two separate contracts. The first was for membership in the Center for Intelligent Information Retrieval (CIIR). This contract was straightforward and below the $25,000 contracting threshold, but nonetheless took several months for approval. The second contract was to cover the staff time of our university collaborators, and this contract took much longer, having not been finally approved until early SEP 00. While we would certainly have been much further along had we not experienced these delays, we had more or less anticipated the situation based on past experiences with government contracting. Staff training and equipment procurement proceeded in parallel with these efforts and is nearly on schedule. Therefore, we do not believe that these contracting delays will prevent completion of the proposed objectives near the proposed timeline. 3. An HP printer capable of printing the color plots needed for graphic visualization of data was ordered and received but has not been working as expected. An ARIEM technician continues to work with HP technical support to remedy the problems.

3. LIFE-CYCLE:

As we began the process of learning how to build a data warehouse and use the data mining tools, we quickly learned that by far and away, the greatest challenge is in the warehousing step itself. The TAIHOD is an immensely complex database, and building its many disparate components into a data warehouse is a very large undertaking. Therefore, much effort still must be expended on this step before we can get to the far more interesting and rewarding step of mining the data. Cost savings in procurement of the software will allow us to get assistance from SAS and from other expert contractors so that we can overcome the somewhat underestimated task involved in this critical step of the process. Although building the data warehouse is not complete, we have begun using data mining to explore the data for important findings in the area of Gulf War Illnesses and motor vehicle crash related injuries per the proposal objectives. We will work with SAS Institute Inc. on two "pilot projects" using SAS Warehouse Administrator and Enterprise Miner software. These pilot projects will involve direct collaboration between TAIHOD personnel and SAS Institute personnel on site at USARIEM. SAS Institute will provide hands-on training for each of these software products tailored to the data from TAIHOD. This will allow us to more fully utilize the software products by efficiently creating a data warehouse, conducting data mining analysis, and producing summary reports for these projects. These projects will also assist us with the efficient creation and implementation of future projects using these software products. In parallel, the collaborators from the University of Massachusetts will proceed according to a statement of work that deviates only slightly from our original proposal.

4. DELIVERABLES:

Our deliverables are the following: 1) Identify demographic and behavioral risk factors for motor vehicle injury among Army personnel, 2) Identify associations between deployment to the Persian Gulf War and subsequent illness, and 3) Explore the data mining capability of the TAIHOD. We are in the process of identifying cases for the motor vehicle study, and will use software developed by the University of Massachusetts for data mining of the text narrative describing the accident. For the Gulf War Illness study, we have identified a cohort for analysis and are beginning to use SAS Enterprise Miner to identify relationships among various variables. There is AMEDD-wide applicability for both of these projects – motor vehicle accidents affect many people and result in large resource usage, and insights into Gulf War Illness can be utilized both to help veterans of the Gulf War and to prevent similar situations for future deployments. A peer-reviewed article describing the background, methods, results, and significance will be written for both deliverables. These reports will be used to measure and report the benefits of the research, and to document project success. One or two methodological manuscripts are expected from our UMA/CS collaborators. It is expected that UMA/CS will take the lead on these papers, which will most likely be submitted to the computer science literature.